Pliable Rejection Sampling

Akram Erraqabi *Montréal Institute of Learning Algorithms, Canada*
Michal Valko *Inria Lille - Nord Europe, France*
Alexandra Carpentier *Institut für Mathematik - U. of Potsdam, Germany*
Odalric-Ambrym Maillard *Inria Saclay - Île-de-France, France*

SequeL – Inria Lille

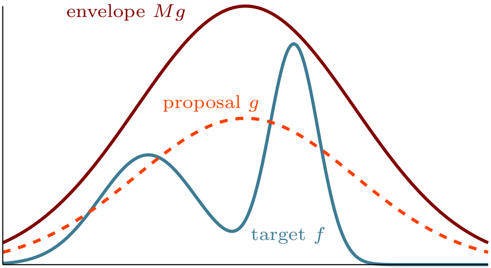
ICML 2016

New York, June 2016

Short Review

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



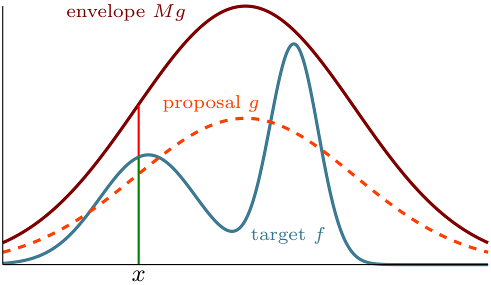
M verifies $f \leq Mg$
Sampling Algo:

1. Sample x from g
2. Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

Short Review

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



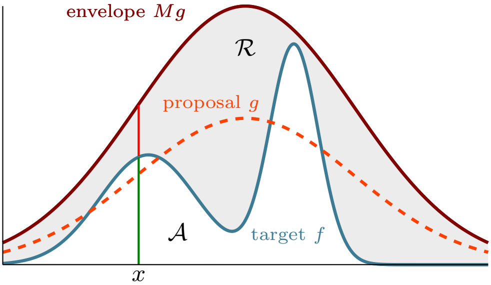
M verifies $f \leq Mg$
Sampling Algo:

1. Sample x from g
2. Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

Short Review

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



M verifies $f \leq Mg$
Sampling Algo:

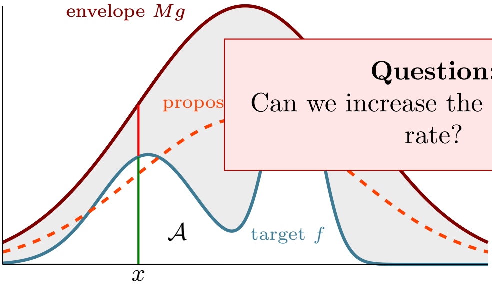
1. Sample x from g
2. Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{M}$$

Short Review

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



Question:
Can we increase the acceptance rate?

M verifies $f \leq Mg$
S... g Algo:

...ple x from g

...ept x as a sample
...n f with
probability $\frac{f(x)}{Mg(x)}$

acceptance rate $= \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{M}$

The setting

Let $d \geq 1$ and let f be a density on $\mathbb{R}^d$.

Goal:

Given a number n of requests to f, what is the number T of samples $Y_1, \ldots, Y_T$ that we can generate such that they are i.i.d. and sampled according to f?

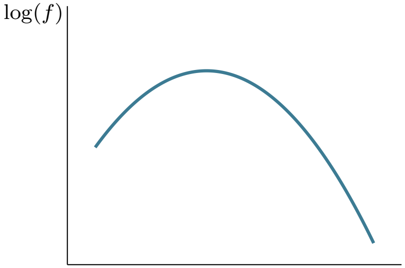
$$\text{acceptance rate} = \frac{T}{n}$$

Can we increase the acceptance rate?

Adaptive Rejection Sampling

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]

- The target f is assumed to be *log-concave* (unimodal)
- The envelope is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$

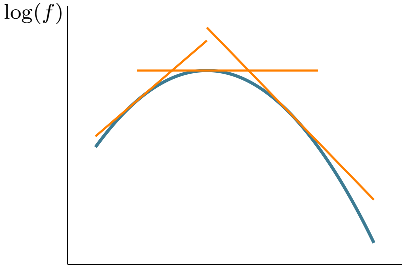


Can we increase the acceptance rate?

Adaptive Rejection Sampling

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]

- The target f is assumed to be *log-concave* (unimodal)
- The envelope is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$

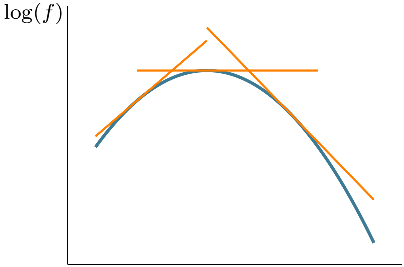


Can we increase the acceptance rate?

Adaptive Rejection Sampling

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]

- The target f is assumed to be ***log-concave*** (unimodal)
- The envelope is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$



Very strong assumption!

Can we increase the acceptance rate?

Improved ARS versions

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks, Best and Tan 1995]

- Can deal with non-log-concave densities.
- Performs a Metropolis-Hastings control for each accepted sample
- At each rejection, the sample is added to $\mathcal{S}$

Convex-Concave Adaptive Rejection Sampling [Gorur and Tuh 2011]

- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)
- At each rejection, the sample is added to $\mathcal{S}$

Can we increase the acceptance rate?

Improved ARS versions

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks, Best and Tan 1995]

- Can deal with non-log-concave densities.
- Performs a Metropolis-Hastings control for each accepted sample
- At each rejection, the sample is added to $\mathcal{S}$

Convex-Concave Adaptive Rejection Sampling [Gorur and Tuh 2011]

- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)
- At each rejection, the sample is added to $\mathcal{S}$

Correlated samples!

Can we increase the acceptance rate?

Improved ARS versions

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks, Best and Tan 1995]

- Can deal with non-log-concave densities.
- Performs a Metropolis-Hastings control for each accepted sample
- At each rejection, the sample is added to $\mathcal{S}$

Correlated samples!

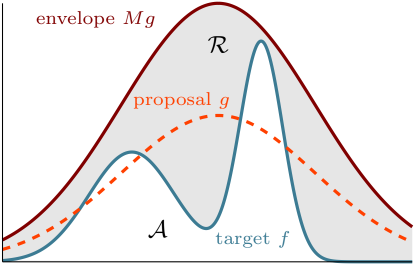
Convex-Concave Adaptive Rejection Sampling [Gorur and Tuh 2011]

- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)
- At each rejection, the sample is added to $\mathcal{S}$

Convexity assumption!

A Pliable Solution

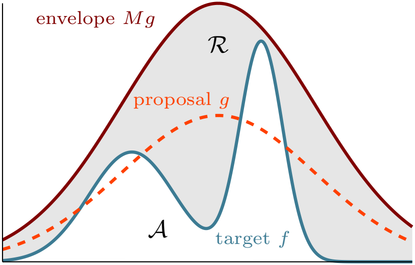
Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

A Pliable Solution

Folding the envelope



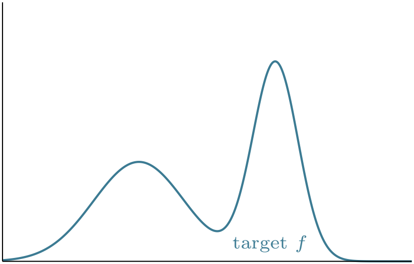
Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

A Pliable Solution

Folding the envelope



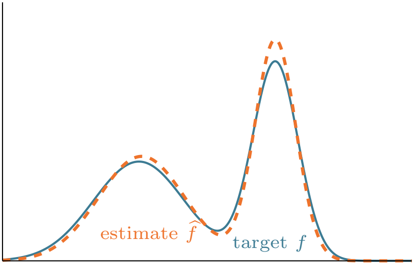
$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

A Pliable Solution

Folding the envelope



acceptance rate $= \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$

Better proposal means smaller rejection area $\mathcal{R}$

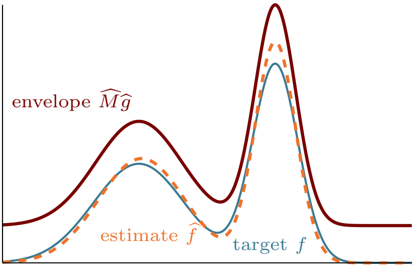
Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$

A Pliable Solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

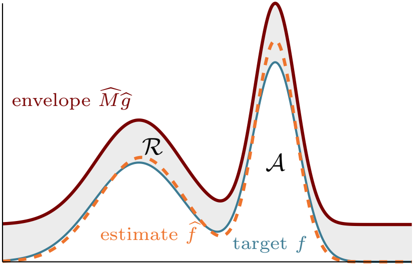
Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

A Pliable Solution

Folding the envelope



acceptance rate $= \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$

Better proposal means smaller rejection area $\mathcal{R}$

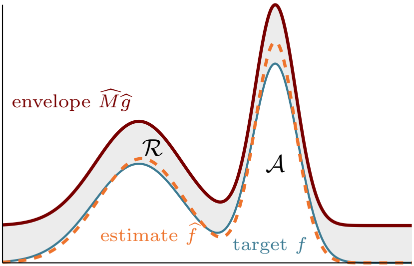
Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

A Pliable Solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

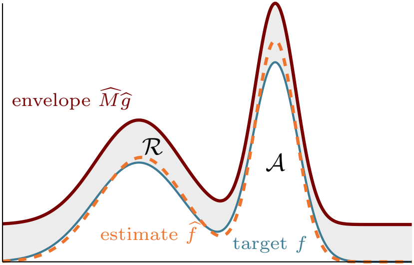
For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

⚠ It should be easy to sample from $\widehat{g}$ ...

A Pliable Solution

Folding the envelope



acceptance rate $= \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

⚠ It should be easy to sample from $\widehat{g}$ ... <u>and $\widehat{f}$!</u>

Visualizing a 2D example

Multimodal case

$$f(x,y) \propto \left(1+\sin\left(4\pi x-\frac{\pi}{2}\right)\right)\left(1+\sin\left(4\pi y-\frac{\pi}{2}\right)\right)$$

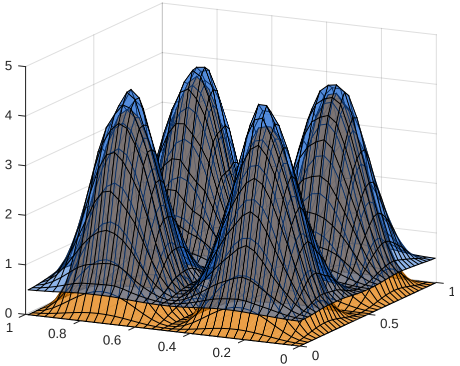


Figure: 2D target density (orange) and the pliable proposal (blue)

Pliable Rejection Sampling

Step 1: Estimating f

- f is defined on $[0, A]^d$, bounded and smooth.
- K is a positive kernel on $\mathbb{R}^d$ (product kernel).
- Let $X_1, \ldots, X_N \sim \mathcal{U}_{[0,A]^d}$. The (modified) kernel regression estimate is

$$\widehat{f}(x) = \frac{A^d}{Nh^d} \sum_{k=1}^{N} f(X_i) K\left(\frac{X_i - x}{h}\right)$$

For an unbounded support density, some extra information is needed to construct a kernel-based estimate.

Pliable Rejection Sampling

Step 1: Estimating f

- f is defined on $[0, A]^d$, bounded and smooth.
- K is a positive kernel on $\mathbb{R}^d$ (product kernel).
- Let $X_1, \ldots, X_N \sim \mathcal{U}_{[0,A]^d}$. The (modified) kernel regression estimate is

$$\widehat{f}(x) = \frac{A^d}{Nh^d} \sum_{k=1}^{N} f(X_i) K\left(\frac{X_i - x}{h}\right)$$

Cost: N requests to f out of n.

For an unbounded support density, some extra information is needed to construct a kernel-based estimate.

Pliable Rejection Sampling

Bounding the gap

Theorem 1

The estimate $\widehat{f}$ is such that with probability larger than $1-\delta$, for any point $x \in [0, A]^d$,

$$\left|\widehat{f}(x) - f(x)\right| \leq H_0 \left(\left(\frac{\log(NAd/\delta)}{N} \right)^{\frac{s}{2s+d}} \right)$$

where H_0 is a constant that depends on the problem parameters.

s is the degree to which f can be expanded as a Taylor expression.

Pliable Rejection Sampling

Bounding the gap

Theorem 1

The estimate $\widehat{f}$ is such that with probability larger than $1-\delta$, for any point $x \in [0, A]^d$,

$$\left|\widehat{f}(x) - f(x)\right| \leq H_0 \left(\left(\frac{\log(NAd/\delta)}{N} \right)^{\frac{s}{2s+d}} \right)$$

where H_0 is a constant that depends on the problem parameters.

s is the degree to which f can be expanded as a Taylor expression.

Remaing Budget: $n - N$.

Pliable Rejection Sampling

Step 2: Generating Samples

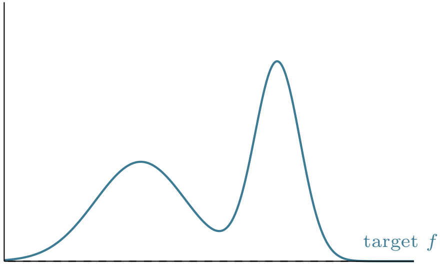
- Remaining requests to f: $n - N$
- Let $r_N = A^d H_C \left(\frac{\log(NAd/\delta)}{N}\right)^{\frac{s}{2s+d}}$
- Construct the *pliable* proposal $\widehat{g}$ out of $\widehat{f}$:

$$\widehat{g} = \frac{\widehat{f} + r_N\, \mathcal{U}_{[0,A]^d}}{\frac{1}{N}\sum_{i=1}^{N} f(X_i) + r_N}$$

- Perform rejection sampling using $\widehat{g}$ and the empirical rejection sampling constant

$$\widehat{M} = \frac{\frac{1}{N}\sum_i f(X_i) + r_N}{\frac{1}{N}\sum_i f(X_i) - 5r_N}$$

The algorithm

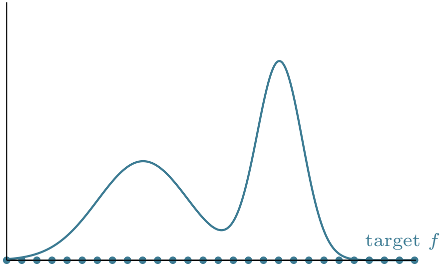


Algorithm: Pliable Rejection Sampling (PRS)

Input: s, n, δ, H_C

Output: $\widehat{n}$ accepted samples

The algorithm



Algorithm: Pliable Rejection Sampling (PRS)

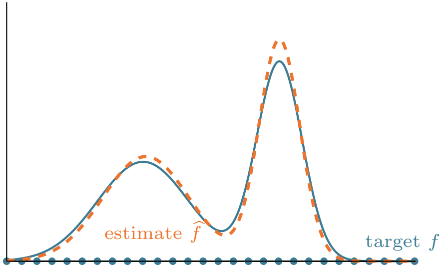
Input: s, n, δ, H_C

Initial Sampling

Draw uniformly at random N samples on $[0, A]^d$

Output: $\widehat{n}$ accepted samples

The algorithm



Algorithm: Pliable Rejection Sampling (PRS)

Input: s, n, δ, H_C

Initial Sampling

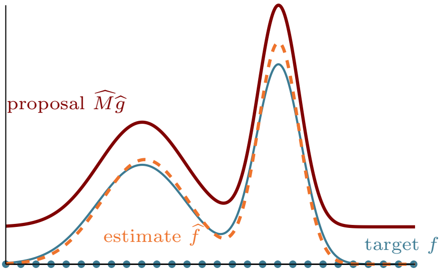
Draw uniformly at random N samples on $[0, A]^d$

Estimation of f

Estimate f using these N samples by kernel regression

Output: $\widehat{n}$ accepted samples

The algorithm



Algorithm: Pliable Rejection Sampling (PRS)

Input: s, n, δ, H_C

Initial Sampling

Draw uniformly at random N samples on $[0, A]^d$

Estimation of f

Estimate f using these N samples by kernel regression

Generating the samples

Sample $n - N$ samples from the *pliable proposal* $\widehat{g}$ and perform Rejection Sampling using $\widehat{M}$ as the envelope constant

Output: $\widehat{n}$ accepted samples

A bound on the acceptance rate

The asymptotic performace

Theorem 2

Under Theorem 1's assumptions and if $H_0 < H_C$, $8r_N \leq \int_{[0,A]^d} f(x)dx$. *Then, for* n *large enough, we have with probability larger than* $1-\delta$ *that*

$$\widehat{n} \geq n \left[1 - \mathcal{O}\left(\frac{\log\left(nAd/\delta\right)}{n} \right)^{\frac{s}{3s+d}} \right].$$

where $\widehat{n}$ *is the number of i.i.d. samples generated by PRS.*

A bound on the acceptance rate

The asymptotic performace

Theorem 2

Under Theorem 1's assumptions and if $H_0 < H_C$ *,* $8r_N \leq \int_{[0,A]^d} f(x)dx$*. Then, for* n *large enough, we have with probability larger than* $1-\delta$ *that*

$$\widehat{n} \geq n\left[1 - \mathcal{O}\left(\frac{\log\left(nAd/\delta\right)}{n}\right)^{\frac{s}{3s+d}}\right].$$

where $\widehat{n}$ *is the number of i.i.d. samples generated by PRS.*

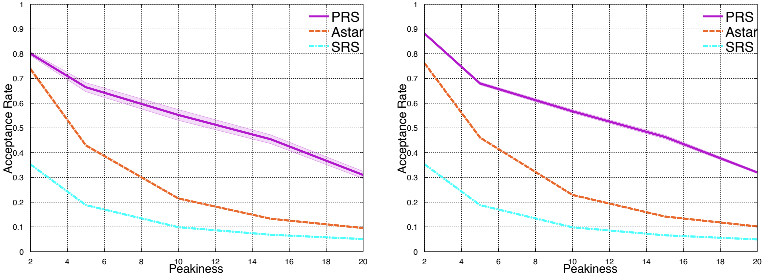
Convergence Rate ↑ with s

Convergence Rate ↓ with d

Experiments

Scaling with peakiness

$$f \propto \frac{e^{-x}}{(1+x)^a},\ a \text{ defines the peakiness level}$$

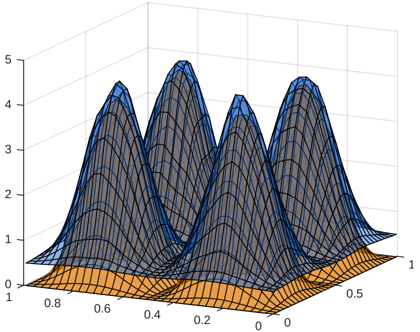


$n = 10^4$ (b) $n = 10^5$

Figure: Acceptance rate *vs.* peakiness

Experiements

Two dimensional example



$n = 10^6$	*acceptance rate*	*standard deviation*
PRS	66.4%	0.45%
$A^\star$ sampling	76.1%	0.80%
SRS	25.0%	0.01%

Table: 2D example: Acceptance rates averaged over 10 trials

Experiments

The Clutter problem

$n = 10^5$, 1D	*acceptance rate*	*standard deviation*
PRS	79.5%	0.2%
A* sampling	89.4%	0.8%
SRS	17.6%	0.1%

$n = 10^5$, 2D	*acceptance rate*	*standard deviation*
PRS	51,0%	0.4%
A* sampling	56.1%	0.5%
SRS	2.10^{-3}%	10^{-5}%

Table: Clutter problem: Acceptance rates averaged over 10 trials

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
    + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

– PRS works only for small and moderate dimensions
    + in favorable cases, it can scale to high dimensions as well
– It does not work well for peaky distributions (posteriors)

Possible extension: Iterative PRS by re-estimating f several times (use the gathered samples to increase its precision)

Thank you!

Questions? feel free to come for a little chat!

Akram Erraqabi
erraqabi@gmail.com
sequel.lille.inria.fr